Statistics 623-423 PROBABILITY IN BIOINFORMATICS AND GENETICS
Spring 2020

TTr 09.25 – 10.40 AM Room DCH 1046

Instructor: Marek Kimmel, Professor, DH 2102, x 5255, kimmel@rice.edu.
TA: Xiaochen Long, xl81@rice.edu

Evaluation
Two tests, midterm (20%) and cumulative final (40%)
Weekly homeworks (40%)
Only for the Stat 623 students, project with a presentation (P/F)
Optional assignments and activity in class (asking questions) may provide extra credit

Honor Code
Homeworks can be discussed in groups. Final redaction has to be done personally. Verbatim copying of homework assignments is a violation.

Tests are under strict Honor Code rules and have to be completed without collaboration or help from others.

Attendance: Essential

Texts:
Ewens and Grant (2005) Statistical methods in bioinformatics (2nd Ed.). Springer (via amazon.com)
Polanski and Kimmel (2007) Bioinformatics. Springer (Available electronically from Fondren).
Rabadan and Blumberg (2019) Topological Data Analysis for Genomics and Evolution. Cambridge

Lecture Notes and Study Guides: The class will be taught using slide presentations with mathematical details worked out on the whiteboard. Study guides will be issued ahead of each test.

Office hours: TA, To be announced.
One-to-one meetings with Prof. Kimmel, by appointment

Contents of the course:
1. Single DNA sequence: Shotgun sequencing, Markov-Chain modeling, patterns and repeats.
2. Multiple DNA or protein sequences: Dynamic programming alignments, substitution matrices and relationships with evolution of structure and function of proteins.
3. Evolutionary Models: Models of nucleotide mutation and substitution, recombination and genetic drift, with applications to genome evolution and population genetics.
4. Stochastic models of evolution and progression of genetic disease.
5. Molecular phylogenetics (tree making): distance matrix, maximum likelihood and parsimony.
6. BLAST: Algorithm for sequence retrieval and comparison.
7. Hidden Markov Models: Viterbi Algorithm and its applications.
8. Systems biology: Analysis and modeling of gene and protein networks

Advances in computing and molecular biology and genomics/genetics, proteomics and other have had an immense impact on the biomedical fields, with broad consequences for humanity. Correspondingly, new areas of probability and statistics have been developed specifically to meet the needs of this area. Probabilistic models are becoming increasingly important in analyzing the huge amount of data being produced by large-scale DNA-sequencing efforts such as the Human Genome Project or individual people's genomes. For example, hidden Markov models are used for analyzing biological sequences, linguistic-grammar-based probabilistic models for identifying RNA secondary structure, and evolutionary models for inferring of sequences from different organisms. This course also describes some of the main statistical applications in the field, including BLAST, gene finding, evolutionary inference and topological data analysis. Only a basic understanding of biological concepts is assumed, and all concepts are explained when used or can be understood from the context.

Course objectives and outcomes:
1. Develop understanding of the importance of chance mechanisms in biological processes on cell and molecule level.
2. Understand how methodologies lead by chance search (such as shotgun sequencing) succeed or fail and how to estimate probability of a correct outcome.
3. Understand how chance drives evolutionary processes.
4. Understand the principles underlying statistical tests in bioinformatics tools such as BLAST.
5. Understand elements of Hidden Markov Models and their use in bioinformatics.

Outcome: Students completing this course will be able to critically evaluate output produced by bioinformatics and genetics/genomic software.

Students with disabilities are encouraged to attend: Regarding accomodations, please contact both the instructor and Disability Support Services at: http://dss.rice.edu/Content.aspx?ekfrm=2147483651 In particular: A student with a disability has a responsibility to contact DSS if any accommodations are required. If accommodations are needed, the student with a documented disability will receive an accommodation letter from DSS to take to each instructor. It is the student's responsibility to deliver these letters as soon as possible. Instructors expect to receive notification early in each semester; students who deliver their letters later may encounter difficulties. Good, early communication is vital to the accommodation process.

Changes: The syllabus is subject to change with reasonable advance notice by the instructor.